Neural Network Models to Identify Medication Relations from Clinical Narratives

Jiazhao Li JIAZHAOL@UMICH.EDU
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI, United States

Jinghui Liu LJINGHUI@UMICH.EDU
School of Information
University of Michigan
Ann Arbor, MI, United States

V.G.Vinod Vydiswaran VGVINODV@UMICH.EDU
Department of Learning Health Sciences and School of Information
University of Michigan
Ann Arbor, MI, United States

Abstract

Electronic health records contain valuable information about patient care and experience, especially related to prescribed medications. Identifying medication relations between drugs and associated attributes automatically from clinical narratives can help develop advanced tools for decision support. In this paper, we investigate the strengths of neural network models to identify eight medication relations. We find that relation extraction is sensitive to complexity of data patterns as well as model capacity. Our results show that bidirectional LSTM models were the best models for the relation classification task, and achieved an overall micro-averaged F1 score of 0.892, outperforming SVM and CNN models, with benefit of bidirectional structure and long dependence memory.

1. Introduction

Electronic health records (EHRs) contain valuable clinical information that can be used for various applications such as clinical decision support, medication reconciliation, public health emergency surveillance, adverse drug events detection, and quality measurements (Demner-Fushman et al., 2009). One of the primary barriers to develop these applications is that much of the information – such as clinical assessment and plans, patient-reported outcomes, effectiveness of treatment, and potential adverse events – are in free text format, and not readily available for downstream processing (Mehrabi et al., 2015). Natural language processing (NLP) can play an important role in discovering clinically-relevant information and making them readily accessible for clinical decision making and other applications.

Consider one such application – identifying adverse drug events from EHRs. Adverse drug events (ADEs), defined as injuries resulting from medical intervention related to prescription drugs (Bates et al., 1995), cause more than 770,000 injuries or death each year.

Data on ADEs observed at the population level provide important evidence regarding the safety of a pharmaceutical product in real-world settings (Sonawane et al., 2018). While significant efforts have been undertaken to identify possible linkages between a drug and a medical condition (Fang et al., 2014; Lee et al., 2018), free text in clinical notes could provide additional evidence of serious adverse events reported by patients.

A variety of approaches have been proposed on identifying medication concepts and relationships from clinical narratives. In this study, we focus on identifying and extracting medical relations associated with drugs from clinical free text narratives. We identify eight attributes associated with a drug, including reason for prescribing the drug, strength of the drug, and any adverse events associated with the drug. While identifying medication concepts by themselves is important, correctly associating mentions of adverse events or reasons with the corresponding mentions of drugs is critical to determine effectiveness of drugs.

Clinical Relevance: The primary purpose of identifying medication relations from clinical narratives is to correctly associate drugs to other concepts related to the medication. This would facilitate linking drugs to ADEs (Li et al., 2018), and more generally, glean structured information from clinical free text, that can enable clinical decision support systems and other applications downstream.

Technical Significance: We formulate the task of relation identification between drugs and its attributes as a separate binary classification task for each attribute-drug pair. In this paper, we investigate traditional feature-rich supervised ML models and deep neural network based models for identifying medication relations. Our investigation highlights merits of incorporating long-range dependencies in identifying medication relations, especially for adverse events and reason for administering the drug.

1.1. Related work

A variety of approaches have been proposed on identifying medication concepts and relationships from clinical narratives. Initial approaches included a combination of rule-based, dictionary-based, and machine learning approaches (Özlem Uzuner et al., 2010; Halgrim et al., 2010; Torii et al., 2011) with varying degree of success on a number of clinical entity-recognition tasks, including medication mention detection and medical problem diagnosis. Patrick and Li (2010) utilized a supervised learning model that incorporated two machine learning algorithms and several rule-based engines to extract medication concepts and relations. Minard et al. (2011) used a hybrid method, with Conditional Random Field (CRF) models for concept extraction and Support Vector Machine (SVM) models for concept extraction, assertion annotation, and relation classification.

In recent years, there has been significant interest in exploring the power of deep neural network architectures for clinical entity and relation extraction tasks. As with traditional approaches, the tasks are set up as sequence labeling problems; however, instead of modeling each token as a hidden variable in a linear chain (bi-directional), the models explicitly ascribe a vector to each node, and pass them through multiple transformations. Wu et al. (2017) developed a recursive neural network (RNN) model trained with word embeddings that achieved a new state-of-the-art performance on the clinical named entity recognition

(NER) task, outperforming the best-reported system that used both manually defined and unsupervised learning features. Yang et al. (2018) developed a recurrent neural networks (RNN) based on long-short-term-memory (LSTM) for medication and ADE identification.

In other works, Yao et al. (2018) used trigger phrases for disease classification, predicting classes with very few examples using trigger phrases and training a convolutional neural network (CNN) with word embeddings derived using the Unified Medical Language System (UMLS). For the relation extraction task, Lv et al. (2016) proposed a CRF model trained on contexts of concepts as baseline and coupled with a deep autoencoder model for relation extraction. In the context of drug safety surveillance, Munkhdalai et al. (2018) explored three models for relation extraction SVM, RNN, and rule-based and found that SVM performed better than the other two. Bidirectional LSTM performed the best among the neural network models. In recent work, Li et al. (2018) focused on extracting ADEs and related information such as medications and indications; using a bidirectional long short-term memory (bi-LSTM) layer, followed by a CRF layer to recognize entities; and a bi-LSTM based attention network to extract relations. They observed that the deep learning model outperformed state of the art systems on MADE 1.0 challenge data. Zeng et al. (2014) exploited a deep convolutional neural network (D-CNN) model on SemEval-2010 Task 8 dataset to extract lexical and sentence level features for classification. The experimental results also demonstrated that the proposed positional features were critical for relation classification. Based on the same dataset, Xu et al. (2016) also proposed a deep recurrent neural (D-RNN) network method using deeper structure to investigate the capacity of integrating information from different abstraction levels.

To facilitate development of novel methods to detect mentions of relations associated with drugs in clinical narratives, Uzuner et al. (2018) developed a shared task as part of the 2018 National NLP Clinical Challenge (n2c2). The relation recognition task was to identify eight kinds of relations between the Drug concept and eight other concept classes related to the Drug, viz. Dosage, Strength, Form, Frequency, Duration, Route, Reason, and ADE. They also provided gold labels for the entities, so that developed systems could evaluate the performance of the relation classifier independently. We use the labeled set provided as part of this shared task to evaluate our approaches.

2. Materials and Methods

2.1. Task description

The goal of the drug relation recognition task is to correctly associate mentions of drugs to mentions of drug attributes in the clinical notes. Specifically, eight attributes related to the administration of the drug are defined, viz. the **Strength** and **Form** of the administered drug, **Route** of administering it, how many (**Dosage**), how often (**Frequency**), and for how long (**Duration**) is it administered, the **Reason** for administering it, and any adverse events (**ADE**) experienced by the patient that can be attributed to the drug.

2.1.1. Data description

The labeled dataset was made available as part of the 2018 n2c2 shared task. It consisted of 505 discharge summaries from the MIMIC-III clinical care database (Johnson et al.,

2016). Concepts and relations relevant to the shared task were annotated by seven domain experts including four physician assistant students and three nurses, following annotation guidelines set by the task organizers. The annotated documents were split into training set (60%, n=303) and test set (40%, n=202).

The counts of the annotated relations in the the two sets are summarized in Table 1. **Form - Drug**, **Strength - Drug**, and **Frequency - Drug** relations are the largest with over $10,000$ instances each, while **Duration - Drug** and **ADE - Drug** relations are the smallest with less than $2,000$ instances each. The train-test split is approximately $60\%-40\%$ for all eight relation classes.

Relation	Train set (n=303)	Test set (n=202)	Total (n=505)
Strength - Drug	6,702	4,244	10,946
Form - Drug	6,654	4,374	11,028
Route - Drug	5,538	3,546	9,084
Dosage - Drug	4,225	2,695	6,920
Frequency - Drug	6,310	4,034	10,344
Duration - Drug	643	426	1,069
Reason - Drug	5,169	3,410	8,579
ADE - Drug	1,107	733	1,840

Table 1: Frequencies of relations annotated in the training and test data sets.

2.2. Overall approach

We modeled the drug-relation identification problem as eight binary classification tasks, one for each attribute, and redefined the task as finding valid associations between the mentions of attributes and the corresponding drugs. The overall approach was as follows: For each of the eight attributes A_i, all (attribute, drug) candidates pairs (e_1, e_2) were constructed, where e_1 is an instance of the attribute A_i and e_2 is a drug. Then, each candidate pair was classified as either "valid" or "invalid", based on contextual features and "valid" pairs were retained as the final associations.

2.2.1. Generating negative instances from labeled data

Following the approach described above, the negative instances were generated from the training data by considering all candidate pairs between drug mentions and any of the eight attributes within the sentence. If the annotations identified a candidate pair as "valid", it was labeled as positive, else it was labeled negative. The candidate pairs were then split into eight classes to create corresponding training sets for the eight relation classifiers.

The distribution of the positive and negative instances in the training set for each binary relation is shown in Table 2. While the class distribution for **Strength - Drug** and **Frequency - Drug** relation tasks are somewhat imbalanced, with three times more negative instances than positive ones, none of the tasks are highly skewed.

Relation	Positive (%)	Negative	Total
Strength - Drug	6,702 (26.4%)	18,660	25,380
Form - Drug	6,654 (42.4%)	9,024	15,678
Route - Drug	5,538 (34.8%)	10,423	15,916
Dosage - Drug	4,225 (42.7%)	5,678	9,903
Frequency - Drug	6,310 (27.1%)	16,975	23,285
Duration - Drug	643 (45.7%)	765	1,408
Reason - Drug	5,169 (48.2%)	5,558	10,727
ADE - Drug	1,107 (59.8%)	743	1,850

Table 2: Distribution of positive/negative samples in train process.

2.2.2. Training relation classification models

For each binary classification task, three supervised classification models were developed: (a) Support Vector Machine (SVM) model, (b) Convolutional Neural Network (CNN) models, and (c) Bidirectional Long Short Term Memory (Bi-LSTM) models. The following sections describe the individual models in more detail.

2.3. Support Vector Machine model

As the first approach, we trained a support vector machine classifier, a traditional feature-rich algorithm, over a set of relation features. We ran cTAKES, a clinical NLP tool (Savova et al., 2010) to perform tokenization, sentence detection, part-of-speech (POS) tagging, chunking, and distance calculation. For each of the eight relation types, candidate relation pairs were identified between instances of the Drug class and all instances of the other class in the relation within a context window of 200 characters from the Drug instance. Since ADE-Drug and Reason-Drug relations were not localized and relatively few, the size of the context window was enlarged to 500 characters.

For each attribute-drug candidate pair (e_1, e_2), the following discriminatory feature classes were extracted: token-level features (form and POS tags) for e_1 and e_2, contextual features (POS tags and entity types of three tokens before and after e_1 and e_2), and distance features (count of tokens between e_1 and e_2 and their entity classes, and whether e_1 and e_2 were in the same sentence). Eight linear-kernel SVM classifiers were trained and tuned using these features to make predictions independently. This model makes use of the traditional feature-rich approach to extract relations in clinical text and serves as the baseline in our experiments.

2.4. Deep Learning Models

We constructed two deep learning based models: convolutional neural network (CNN) and bidirectional long short term memory (Bi-LSTM). The architectures are shown in Fig. 1 and Fig. 2. The same embedding layer was shared by two models introduced in (Sec. 2.4.1) following by CNN (Sec. 2.4.2) and Bi-LSTM (Sec. 2.4.3).

2.4.1. Embedding layers

We processed the corpus using Stanford CoreNLP (Manning et al., 2014), an open source natural language process software, for pre-processing the corpus including sentence and token boundary detection, part-of-speech (POS) tagging, and named entity recognition (NER). For each type of relation, samples were generated with potential entity pair in the center of fixed length of token list. Tokens were embedded with four features: pre-trained word-level token embedding, POS tags, bi-directional relative position of target entities and NER tags, shown in the left of Fig.1. Sentence and token boundaries as well as POS tags were detected using Stanford CoreNLP, and NER tags were looked up from annotation training files. In order to combine with lexical-semantic knowledge in health-care domain, word embedding are pre-trained on MIMIC-III using publicly available *word2vec* model (CBOW) (more details in Sec. 3.2).

Given a sentence $S = (w_1, w_2, \ldots, w_n)$ with target entity $e_1 = w_{t1}$ and $e_2 = w_{t2}$, we first transformed each token into a real-valued vector to offer lexical-semantic feature with pre-trained model introduced above. POS tags and NER tags of tokens were also embedded with one-hot encoder to serve as the supplement of token types. To additionally capture information about the position of the target entities, we incorporated bi-directional word relative position embeddings: the sequential numeric with target entity as centric number 0. For the given sentence in Fig.1, the **Reason** entity e_1 (consisted of two tokens *Graved* and *disease*) and **Drug** entity (consisted of tow tokens : *radioactive* and *iodine*) were all embedded with 0 as the center of sequence.

Each sentence can be represented as a matrix W_V. Let $\boldsymbol{x}_i \in \mathbb{R}^k$ be the k-dimensional word vector corresponding to the ith word in the sentence, where each vector was a concatenation of the word embedding and one-hot encoding over POS tag, NER tag and bi-directional relative position. Consequently, $\boldsymbol{x}_{i:i+j}$ can represent a set of words $\boldsymbol{x}_i, \boldsymbol{x}_{i+1}, ...\boldsymbol{x}_{i+j}$. A sentence of length n (padded or segmented where necessary) can be represented as a train patch $\boldsymbol{x}_{1:n} \in \mathbb{R}^{n \times k}$.

2.4.2. Convolutional neural network

We constructed eight relation classifiers using a convolutional neural network (CNN) model based on the architecture proposed by Kim (2014) consisting of N-gram convolutional layers and fully-connected inference layer.

N-gram convolutional layers: Inspired by N-gram features in a traditional NER model, a convolutional layer can learn to recognize short phrase patterns by applying various sizes of convolutional kernels to a sliding window of size N. For instance, if we denote a bigram kernel as $\boldsymbol{w} \in \mathbb{R}^{N \times k}$, where $N = 2$ on a bigram $\boldsymbol{x}_{i:i+1}$, a phrase-level feature c_i is given by:

$$c_i = f(\boldsymbol{w} \cdot \boldsymbol{x}_{i:i+1} + b) \tag{1}$$

where b is a bias term and f is the activation function that maps each N-gram context window around a word to a scalar feature. In order to learn diverse features at phrase-level, we applied M kernels for each phrase size (N-gram). Under single kernel projection, one

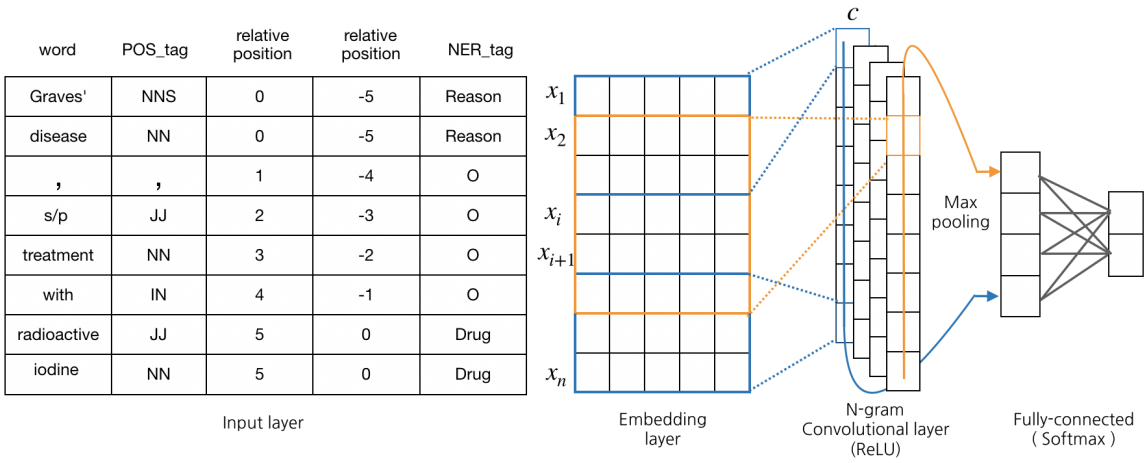

Figure 1: Schematic diagram of the convolutional neural network model for relation recognition task.

sentence with length of l can be embedded into one feature vector $\boldsymbol{c}$:

$$\begin{aligned} \boldsymbol{c} &= [c_1, c_2, c_3, \ldots, c_l] \\ \boldsymbol{C} &= [\boldsymbol{c_1}, \boldsymbol{c_2}, \ldots, \boldsymbol{c_M}] \end{aligned} \tag{2}$$

with $\boldsymbol{c} \in \mathbb{R}^l$, $\boldsymbol{C} \in \mathbb{R}^{l \times M}$.

Max-pooling layer: A Max-pooling layer was applied on the sentence vector $\boldsymbol{C}$ in order to determine the importance of individual windows. For each task, the probability of relation can be achieved by passing through a fully-connected layer with activation function *Softmax*.

$$\alpha = Softmax(\boldsymbol{w}^T \cdot \boldsymbol{C} + b) \tag{3}$$

where $\boldsymbol{w} \in \mathbb{R}^{l \times M}$ represent the weights learned by the model.

2.4.3. Bidirectional Long Short Term Memory model

We also constructed eight relation classifiers based on recurrent neural network: Bidirectional Long Short-Term Memory architecture firstly proposed by Baldi et al. (1999). The center idea of using this model was to exploit the benefit of recurrent neural network learning long range dependence between target tokens in the sentence, rather than short range linguistic information from *n-gram* kernels by convolutional nerual network. Compared with classic recurrent neural network, LSTM memory cell overcame vanishing gradient problem. The unfolded architecture is shown in Fig. 2, consisting of three components: an embedding layer, two Bi-LSTM layers and fully-connected *Softmax* output layer.

Embedding layer processed input sequential sentence into a $n \times k$ matrix patch $x_{i:n}$, which is the same as CNN model (Sec.2.4.1). Single Bi-LSTM layer had two directions of propagation: forward and backward. In each direction, recurrent memory cell in first Bi-LSTM layer took in one token embedding x_i each time as input and computed the hidden

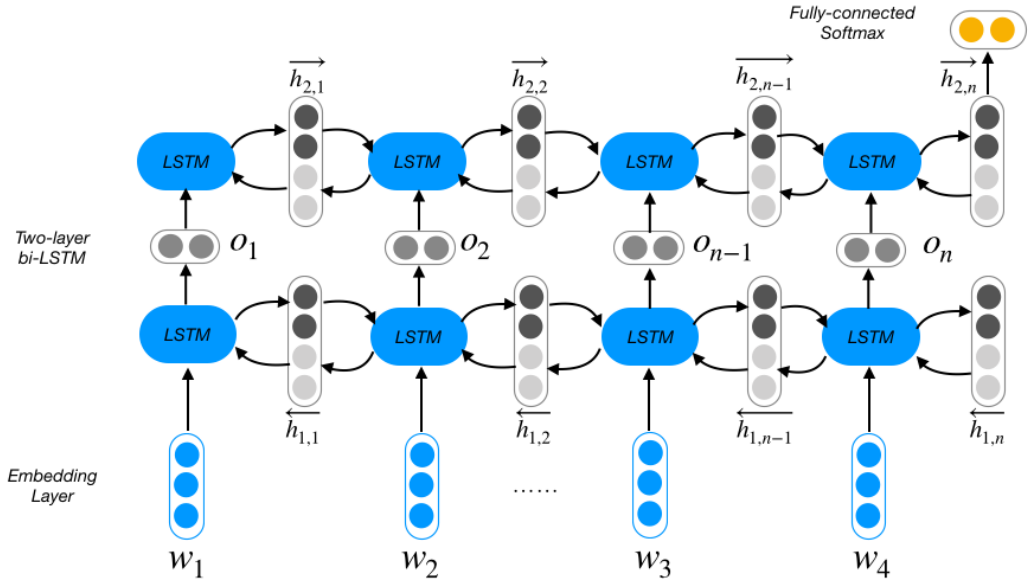

Figure 2: Schematic diagram of the Bi-LSTM model for relation recognition task.

vector for each direction $\overrightarrow{h_{i,j}}, \overleftarrow{h_{i,j}} \in \mathbb{R}^m$, where m is the length of hidden units, and output vector o_i with the same size of hidden vector. At time-step t, the memory cell was updated with the following function (Graves and Schmidhuber, 2005):

$$
\begin{aligned}
i_t &= \sigma(W_{x_i} x_t + W_{h_i} h_{t-1} + W_{c_i} c_{t-1} + b_i) \\
f_t &= \sigma(W_{x_f} x_t + W_{h_f} h_{t-1} + W_{c_f} c_{t-1} + b_f) \\
c_t &= f_r c_{t-1} + i_t tanh(W_{x_c} x_t + W_{h_c} h_{t-1} + b_c) \\
o_t &= \sigma(W_{x_f} x_t + W_{h_f} h_{t-1} + W_{c_f} c_{t-1} + b_f) \\
h_t &= o_t tanh(c_t)
\end{aligned}
\tag{4}
$$

where σ is the sigmoid activate function, and i, f, o and c are resprectively the *input gate*, *forget gate*, *output gate*, and *memory cell*.

The second Bi-LSTM layer memory cell took the concatenation of forward and backward output vector from first layer as input to extract higher level patterns features. The hidden vector $h_{2,n}$ at the end of second Bi-LSTM layer was followed by a fully-connected layer with *Softmax* as the activation function and dropout layer in order to produce the binary classification prediction and prevent overfitting.

3. Experiment and Results

We conducted our experiments on three models: SVM, CNN and Bi-LSTM over eight relation extraction tasks with n2c2 dataset.

3.1. Feature selection

Selection and generation of feature sets, shown in Table 3 were introduced in Sec.2.3 for SVM model and Sec.2.4.1 for deep learning models.

Classifiers	Fearture sets
SVM	syntactic pattern, POS tag, NER tag, neighbor tokens, sentence distance, relative position
CNN & Bi-LSTM	word embedding, POS tag, NER tag, relative position

Table 3: Models and corresponding feature sets

3.2. Word embedding in health-care domain

In order to combine with lexical-semantic knowledge in health-care domain, 50K clinical notes from MIMIC-III dataset were used as corpus to learn tokens embedding on public available *word2vec* model (Mikolov et al., 2013). Some settings of models are followed: window size of 5, negative sampling, lower case and CBOW model. In order to account for unknown tokens during the test phase, we used a random embedding vector in Gaussian distribution as replacement.

3.3. Specification and Training of Deep learning models

We implemented both deep learning neural network using Keras (Chollet et al., 2015). The embedding of each token had dimensions of 156, including 100 from word embedding and 56 from other discrete features. For CNN model, 200 kernels were used for each ngram type (from 1×156 to 5×156) to capture different types of patterns in convolutional layers. For Bi-LSTM model, two layers of Bi-LSTM were used with 30 hidden units in each memory cells. In the binary classification task, we updated the parameters with respect to the loss function: *CrossEntropy* between target and prediction as shown below:

$$\mathscr{L} = -ylog(\alpha) + (1-y)log(1-\alpha) \tag{5}$$

where y is the ground truth of relation from 0 to 1 and α is the prediction probability. We trained both two neural networks with Adam optimizer (Kingma and Ba, 2014) with the mini-batch size of 25. The learning rate was set to 0.001. We kept 12.5% of the training corpus for a validation corpus and we implemented early stopping with a patience of 10 epochs without performance improvement. Finally, a rate of 0.5 dropout layer with was applied in input embedding and fully-connected block to avoid overfitting. Other matrices were initialized with random values following a Gaussian distribution. The fine-tuned hyperparameters from validation corpus were given in Table 4. We reported F1-score of test corpus in Table 5 computed with the official evaluation script provided during the 2018 n2c2 Shared Task challenges.

Hyperparameters	CNN	Bi-LSTM
Sentence length	50	50
Word embedding dim	100	100
Feature embedding dim	56	56
Kernel height	1-5	-
Number of kernels	200	-
Hidden layer units	-	30
Dropout probability	0.5	0.7
Validation set ratio	0.125	0.125
Number of epochs	20	30

Table 4: Hyperparameters used for training the neural network models.

3.4. Results

For each model: SVM, CNN or Bi-LSTM and each relation type, we reported F1-score computed with the official evaluation script provided during the 2018 n2c2 Shared Task challenges. Results of the experiments are presented in Table 5. Bi-LSTM achieved the best in overall metric among three models. ADE-drug (F1 of 0.701) and Reason-Drug (F1 of 0.721) relations are the most difficult types to make the prediction with the reason of long distance between two target entities. Because of lack of enough train samples, Duration-Drug (F1 = 80.20) type comes to third worst as well as ADE-drug mentioned. In contrast, with short dependence and enough train samples, the other relation types are all around F1 of 0.9 (from 0.897 to 0.952).

3.4.1. Comparison between SVM and Deep Learning Models

SVM is traditional supervised discriminative classifier chosen as the baseline of tasks with overall (micro) performance (F1 of 0.656) compared with two deep learning models: CNN (F1 of 0.820) and Bi-LSTM (F1 of 0.892). Two deep learning models significantly outperformed SVM on most relations extraction tasks with the benefit of higher capacity of models themselves on overall F1 score metric.

3.4.2. Comparison between CNN and Bi-LSTM

The advantage of recurrent neural networks on sequence data is apparent compared with short dependence from $N - gram$ kernels, with all tasks higher than convolutional neural network. The largest improvement is Duration-Drug (with 0.1324 increase in F1), followed by ADE-Drug (with 0.1221 increase in). Our model did not distinguish between intra- and inter- sentence relations, but instead considered that related entities had to occur within a window of 50 tokens.

4. Discussion

With the benefits of bi-direction and long dependence mechanism of the long short term memory, we are able to efficiently train a highly performing deep-learning algorithm to

Relation type	SVM	CNN	Bi-LSTM
Strength - Drug	**0.9355**	0.8495	0.9333
Form - Drug	0.6477	0.9142	**0.9519**
Route - Drug	0.7444	0.8600	**0.9192**
Dosage - Drug	0.6120	0.8495	**0.9355**
Frequency - Drug	0.4839	0.8254	**0.8973**
Duration - Drug	0.3307	0.6696	**0.8020**
Reason - Drug	0.3570	0.6623	**0.7208**
ADE - Drug	0.4431	0.5779	**0.7007**
Overall (micro)	0.6558	0.8203	**0.8920**
Overall (macro)	0.6565	0.8109	**0.8890**

Table 5: Performance results of three models on the eight relation classification tasks, as measured by the Lenient F1 scores.

extract medication relations from medical texts. The results show the efficacy of using this method in comparison with traditional machine learning algorithm. The results show that deep learning models perform well on entities that have more variations in its form and length, such as Duration, Reason, and ADE. For capturing short entities with simple forms, like Strength, it turns out that SVM can have high accuracy as well, and it fails to perform consistently on other entities. With the capability to detect longer sequence, deep learning models also show a more balanced score in the eight classes.

4.1. Limitations

Since the clinical free text data has high variation in its vocabulary and structure, different strategies of preprocessing the raw data can directly affect the performance of the algorithms. Choices on the tokenization of the text and the width of window to filter potential candidates will change the way how algorithms learn pattern from the data. It is possible that some of the preprocessing steps have more influence on the final performance of the system than the changing of algorithm. In feature engineering of deep learning model, we didn't consider the indicator of whether two target entities intra- or inter- the same sentences, which probably could decrease the noise from other tokens. Also, considering the popularity and power of attention mechanism in NLP area, the stack of Bi-LSTM and attentive structure could be expected to issue this problem having some improvement.

4.2. Error Analysis

Large diversity on results with the same model between eight classification tasks shows sensitivity of models and are highly related with difficulty of tasks. Models performed better on Strength-Drug, Dosage-Drug and Frequency-Drug tasks due to relatively simple pattern of the instances: short distance between Drug name and ADE (Strength: digit +

units, Dosage: digit, Frequency: abbreviation) such as:

Strength – Drug : ***lisinopril*** *was decreased to* ***20mg*** *po daily*

Dosage – Drug : ***multivitamin*** *Tablet Sig:* ***One (1)*** *Tablet PO once a day.*

Frequency – Drug : ***oxycodone*** *5mg* ***q4h p.r.n***

For tasks with complex patterns, models showed its limitations: difficulty on dealing with the long dependence between Drug and ADE/Reason which could appear in different sentences. Though wider and deeper Bi-LSTM is powerful enough to capture this long dependence, complexity of model and small train dataset resulted in overfitting problem.

5. Conclusion

The 2018 n2c2 shared tasks on adverse drug event and medication extraction from clinical narratives provided a rich test bed to evaluate deep neural network-based models. In our experiments, the bidirectional long short term memory model outperformed convolutional neural network and SVM baseline on the relation recognition task. Additional experiments are planned to further investigate the role of health-care domain pre-trained token embedding representations in neural network models, and developing a joint model for both concept and relation recognition tasks.

References

Pierre Baldi, Søren Brunak, Paolo Frasconi, Giovanni Soda, and Gianluca Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11):937, 1999.

David W Bates, David J Cullen, Nan Laird, Laura A Petersen, Stephen D Small, Deborah Servi, Glenn Laffel, Bobbie J Sweitzer, Brian F Shea, Robert Hallisey, et al. Incidence of Adverse Drug Events and Potential Adverse Drug Events: Implications for Prevention. *JAMA*, 274(1):29–34, 1995.

François Chollet et al. Keras. https://keras.io, 2015.

Dina Demner-Fushman, Wendy Chapman, and Clement Mcdonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42:760–72, 09 2009. doi: 10.1016/j.jbi.2009.08.007.

Hong Fang, Zhenqiang Su, Yuping Wang, Aaron Miller, Zhichao Liu, Paul Howard, Weida Tong, and Simon M Lin. Exploring the fda adverse event reporting system (faers) to generate hypotheses for disease monitoring. *Clinical Pharmacology and Therapeutics*, 95 (5):496–498, May 2014. doi: 10.1038/clpt.2014.17.

Alex Graves and Jrgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

Scott Halgrim, Fei Xia, Imre Solti, Eithon Cadag, and Özlem Uzuner. Extracting medication information from discharge summaries. In *Proceedings of NAACL-HLT 2010 Second LOUHI Workshop on Text and Data Mining of Health Documents*, pages 61–67, 2010.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Andy Jinseok Lee, Sunyang Fu, and V.G.Vinod Vydiswaran. Supervised learning approach to link prediction in fda adverse event reporting system (faers) database network. In *Proceedings of the AMIA Annual Symposium*, 2018.

Fei Li, Weisong Liu, and Hong Yu. Extraction of information related to adverse drug events from electronic health record notes: Design of an end-to-end model based on deep learning. *JMIR Medical Informatics*, 6(4):e12159, 2018.

Xinbo Lv, Yi Guan, Jinfeng Yang, and Jiawei Wu. Clinical relation extraction with deep learning. *International Journal of Hybrid Information Technology*, 9(7):237–248, 2016.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M. Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul R. Dexter, C. Max Schmidt, Hongfang Liu, and Mathew J. Palakal. Deepen: A negation detection system for clinical text incorporating dependency relation into negex. *Journal of Biomedical Informatics*, 54:213–219, 2015.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Anne-Lyse Minard, Anne-Laure Ligozat, Asma Ben Abacha, Delphine Bernhard, Bruno Cartoni, Louise Deléger, Brigitte Grau, Sophie Rosset, Pierre Zweigenbaum, and Cyril Grouin. Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. *Journal of the American Medical Informatics Association*, 18(5):588593, 2011.

Tsendsuren Munkhdalai, Feifan Liu, and Hong Yu. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: Classical learning versus deep learning. *JMIR Public Health Surveillance*, 4(2):e29, 2018.

Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of American Medical Informatics Association*, 17(5):514–518, Sep–Oct 2010.

Jon Patrick and Min Li. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of American Medical Informatics Association*, 17(5):524527, Sep–Oct 2010.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

Kalyani B. Sonawane, Ning Cheng, and Richard A. Hansen. Serious adverse drug events reported to the fda: Analysis of the fda adverse event reporting system 2006-2014 database. *Journal of Managed Care & Specialty Pharmacy*, 24(7):682–690, 2018. doi: 10.18553/jmcp.2018.24.7.682. URL https://doi.org/10.18553/jmcp.2018.24.7.682. PMID: 29952714.

Manabu Torii, Kavishwar B. Wagholikar, and Hongfang Liu. Using machine learning for concept extraction on clinical documents from multiple data sources. *Journal of American Medical Informatics Association*, 18(5):580–587, 2011.

Özlem Uzuner, Ma Filannino, and Amber Stubbs. Adverse drug events and medication extraction in ehrs: Track 2 of the 2018 n2c2 shared-task and workshop, 2018. URL https://n2c2.dbmi.hms.harvard.edu/track2.

Yonghui Wu, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu. Clinical named entity recognition using deep learning models. In *Proceedings of AMIA Annual Symposium*, page 18121819, 2017.

Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. Improved relation classification by deep recurrent neural networks with data augmentation. *arXiv preprint arXiv:1601.03651*, 2016.

Xi Yang, Jiang Bian, and Yonghui Wu. Detecting medications and adverse drug events in clinical notes using recurrent neural networks. In *Proceedings of Machine Learning Research: 1st International Workshop on Medication and Adverse Drug Event Detection*, pages 1–6, 2018.

Liang Yao, Chengsheng Mao, and Yuan Luo. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. In *Proceedings of IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*, pages 70–71, 2018.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C14-1220.